# Critical Issues in Bioinformatics and Computing

*by Someswar Kesh, PhD, and Wullianallur Raghupathi, PhD*

## Abstract

This article provides an overview of the field of bioinformatics and its implications for the various participants. Next-generation issues facing developers (programmers), users (molecular biologists), and the general public (patients) who would benefit from the potential applications are identified. The goal is to create awareness and debate on the opportunities (such as career paths) and the challenges such as privacy that arise. A triad model of the participants' roles and responsibilities is presented along with the identification of the challenges and possible solutions.

## Introduction

"Bioinformatics" is defined by the National Institutes of Health as the "research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data."[1] The exponential growth in the amount of such data has necessitated the use of computers for information cataloging and retrieval, while a more global perspective in the quest for new insights into health and disease and the resulting data mining also underscore the need for bioinformatics.[2]

Both sophisticated hardware and complex software play an increasingly critical role in the analysis of genomic data, and the accelerated maturation of the field of bioinformatics has implications for computing and life sciences professionals as well as the general public. Like computational biology, bioinformatics is anchored in the life sciences as well as computer and information sciences and technologies. Its interdisciplinary and integrative approach draws from fields such as mathematics, physics, computer science and engineering, biology, and behavioral science. The generally accepted subdisciplines include (1) development of new algorithms and statistics with which to assess relationships among members of large data sets; (2) analyses and interpretation of various types of sequences, domains, and structures; and (3) development and implementation of tools that enable efficient access and management of different types of information.[3–5]

Recently, the field of bioinformatics has experienced rapid growth. However, as with other young disciplines, it now faces a host of critical issues. Successfully addressing these key issues is essential to further progress in the field.

This article develops and applies a triad model to discuss current and future issues in bioinformatics from the perspective of computing professionals. A working model of the participants' roles and responsibilities is presented, along with the identification of the challenges and possible solutions. This model can also be used in the future as newer issues emerge. The goal is to create awareness and debate on the opportunities (for example, career paths) and challenges (for example, privacy) that arise. Furthermore, the ideas discussed here can be operationalized with future empirical research.

## The Triad Model

Figure 1 describes the model for understanding the roles and relationships among the various participants, including computing professionals, users, and the general public. The roles and responsibilities are by no means mutually exclusive. On the contrary, considerable overlap exists among the three.

### *Participants*

Computing professionals, including developers, programmers, consultants, and vendors, should be concerned with building and testing robust applications and performance issues such as correctness of data, reliability, and real-time processing, and integration and management of data deployed to serve multiple purposes simultaneously.

Users, including molecular biologists and other scientists in the life sciences, are concerned with data input and user interfaces, analysis and analytical tools, and interpretation in a shared, global environment. Not only are new genetic tests and procedures required, but clinical information must also be integrated and analyzed continuously to facilitate identification of adverse reactions. Furthermore, the ability to immediately detect promising new drug applications as well as bring new products rapidly to the market are significant challenges. At this time, users are constrained by incomplete, piecemeal tools with poor usability. A primary objective is to work with the other groups in developing sophisticated applications and analytical tools that facilitate fast querying and data mining. The sharing of resources without reinventing the wheel is another challenge.

The public at large is concerned with implications of potential medical applications, ethics, privacy, potential misuse of data, and public and social policies. Almost every citizen is involved in these issues, including social workers, legal and medical professionals, lawmakers, patients, and other participants, including pharmaceutical companies and healthcare providers.

### *Roles and Responsibilities*

The intersecting area in Figure 1 depicts overlapping roles and responsibilities of participants in the application of the triad model. For example, the public should decide what can and should be ethical and legal. This will directly place limits on the type of research the user may perform. The user, on the other hand, can and should join this public debate. Once it has been decided *what* can be done, the relationship between the user and the computing professional comes into play to determine *how* computing technology

can assist the user. This is not a static relationship, but rather a dynamic equilibrium, where the participants in the model will have to decide on the point of equilibrium at a particular point in time and in a specific social context.

The ongoing debate on Iceland's medical database is an example of how medical information-related issues have become significant. A law passed in the late 1990s enabled creation of a medical database that includes medical and family history records as well as genetic information on all Icelanders. The database was contracted out to a third-party biotechnology firm. Questions have been raised regarding violations of medical and personal privacy; medical stereotyping of individuals, families, or the entire population; potential discrimination based on medical or genetic data; and a monopoly on medical research and drugs by large companies. Iceland's medical database issues may affect policies in other parts of the world. In an online article, "For Sale: Iceland's Genetic History," Oksana Hlodan identifies the following issues: [6]

- Who has the right to access and use our personal genetic data?
- Who controls the data?
- If medical records are used as a community resource, should they not be available to all research facilities within the community?
- Will the medication for a disease discovered through population genetics studies be available to the participants?
- Can anybody own pieces of our genome through patents, copyrights, and so on?
- Should genetic testing be done, and how scientifically reliable is it?
- How will other citizens perceive an individual whose genetic tests reveal a potential disease?
- Will the data lead to discrimination?

The triad model described here can be generalized for the larger field of health information management (HIM), which encompasses all aspects of the healthcare industry, including the flow of information therein. The participants would include patients, healthcare providers (including physicians, nurses, health maintenance organizations [HMOs], insurance companies, hospitals, pharmacies, and medical testing agencies), and federal programs such as Medicare. The gathering, storage, processing, and dissemination of the disparate and complex medical information generated by the overlapping interaction between these entities will result in the need to address privacy and security issues. The dynamics of the interaction and the resultant outcomes can be studied using the triad model.

Table 1 summarizes the roles and responsibilities of the participants in the triad model in greater detail. It must be noted these can also be construed as the challenges and tasks faced by the discipline as a whole. The challenges faced include gene discovery and analysis, and issues in the potential revelation of previously unknown relationships with respect to genetic structure and function. This becomes particularly critical in light of the vast amount of data being produced by the Human Genome Project.

## New Challenges for Computing Professionals

The scientific community has marked a significant milestone in the study of genes, the completion of the "working draft" of the human genome. This work, which was recorded in special issues of the journals *Nature* and *Science* in 2001, heralds a new beginning for advances in the prevention, diagnosis, and treatment of many genetic and genomic disorders. The availability of this wealth of raw data has a significant effect on the field of bioinformatics, with a great deal of effort being spent on effectively and efficiently storing and accessing these data, as well as on new methods aimed at mining the data in order

to make revolutionary medical discoveries.[7] These advances have generated numerous new and exciting challenges with which computing professionals will have to grapple.

A large variety of genomic data sources have emerged, resulting in inconsistent terminology and data formats. Many of these come from independent studies that were organism based, such as for cancer research. While much of this information is publicly available over the Internet, comparison and unification are critical for much of the sequence analysis that remains to be done. However, the fact that these data sources were developed for different purposes by different researchers using different methods often makes the data difficult to unify. Regarding data standards, the emergence of the macromolecular crystallographic information file (mmCIF) and extensible markup language (XML) provides standards that can produce a common format for data. It is critical that the bioinformatics community either decide on or gravitate toward one common format that will make data sharing vastly easier.

### *An Integrative Framework*

Additionally, collaborative research requires conceptualization and implementation of an integrative framework. Apart from standardization of data formats, this will require development of Web-based user interfaces, standards for access to the data and data warehousing capabilities, as well as interoperable software components. The development of a standardized, Web-based, globally distributed view is critical in the light of researchers working together across several languages and countries. A standardized interface to the multiple heterogeneous databases is an important objective for developers.

Two distinct approaches have been used for data warehousing. IBM uses a *federated database,* in which the data remain in the original separate sources and are accessible with a single query. The data from various sources are brought into a *data warehouse,* where data freshness depends on the frequency of data replication. The issue of which approach is more useful and when is yet to be determined.

Examples of data sources for a federated database or data warehouse are the three primary sequence databases: GenBank (NCBJ), Nucleotide Sequence Database (EMBC), and the DNA Databank of Japan (DDBJ). These are repositories for raw sequence data, but each entry is extensively annotated and has a features table to highlight the important prospects of each sequence. The three databases exchange data on a daily basis.[8]

Interoperability among software components is a crucial goal for successful collaborative work. Object management groups (OMG) and a life sciences research domain task force's goal to establish common object request broker architecture (CORBA) as the standard for interoperable software components offer potential.[9]

### *Future Computing Needs*

While the knowledge gained from the sequencing of the human genome via bioinformatics is expected to change our lives, more powerful and robust computing is needed to develop the tools for genetically based drug design, medical diagnosis and treatment, and agricultural application, among others. The power and robustness should come from development of both software algorithms and hardware. Many traditional algorithms, including Bayesian statistics, dynamic programming, and Markov chains, have already been used for sequencing.

With the enormous size of databases today, the efficiency of these algorithms is critical for successful use. Dynamic programming, for example, can considerably slow down in multiple sequence alignments because the complexity of the calculations increases for more than two sequences. However, improvements in the algorithms and use of heuristics have improved the situation significantly. Future research should focus on development of such heuristics.[10]

Moreover, mining the data for patterns is essential for newer discoveries. Pattern recognition algorithms and neural networks have been applied to bioinformatics research. Neural networks can also be applied to classification as well as decision problems.[11] Other artificial intelligence-based algorithms, like case-based reasoning (CBR), can be useful in this regard.

The issue is to embellish the currently available algorithms and heuristics as well as develop new ones to deal with the need for sequencing, prediction, and pattern recognition. Comparative studies of the effectiveness and efficiency of these algorithms are essential for further applications.

The term “deep computing” for bioinformatics research, implies the use of powerful machines executing sophisticated software based on innovative algorithms to solve complex problems like mapping, modeling, and visualization. From a hardware perspective, both a supercomputing approach and a distributed computing approach have been used in bioinformatics. Grid computing allows geographically distributed organizations to share applications data and computing resources.[12] While the distributed approach is less expensive, it raises further issues endemic to distributed processing and data distribution, particularly those over Internet services.

To facilitate access, several tools have been developed or are works in progress. These tools include GeneX, an example of a system that helps with the storage, organized retrieval, and analysis of gene expression data. Among the most important software tools for the understanding of DNA and protein sequences are sequence similarity and alignment tools such as Basic Local Alignment Search Tool (BLAST) and a sequence alignment algorithm using a flat file format known as FASTA. Figure 2 is a sample screen capture of the BLAST interface. One can visualize the complexity of the back-end databases and the front-end query tools with which BLAST deals. These tools allow one to compose an unknown sequence with a database of sequences from other organisms that are better understood. These programs report the hit in the database, along with the estimated statistical significance of the hit.[13]

DiscoveryLink is described as a middleware software product from IBM. It can be used to build a federated database application. A prototype system called MyGrid is being developed at universities in the UK. The new system will allow biologists to analyze information in many databases in a standardized fashion, which until now required many types of custom-built software. It is reported that with MyGrid, biologists will not become programmers, for the team is using software agents to help translate and standardize the contents of conflicting formats. MyGrid should automatically find any information relevant to the study, searching for genomic and proteomic data, regulatory networks, and any other relevant facts.[14]

The robustness of data submitted to the primary database is important in the context of bioinformatics software. Much of the progress in bioinformatics is in fact due to the accelerated rate at which sequence data are being produced. Bioinformatics is required at several different stages during DNA sequencing. First, the data produced at every stage of generation and analysis must be captured in real time. Second, sophisticated software algorithms are required to assemble, edit, and compare the sequence data. Genomic databases need to facilitate the storage and analysis of large amounts of data, but also have a user-friendly format and graphical display to allow relevant data to be displayed and analyzed.

Beyond storage and integration, the computing capabilities required for these new scientific developments are diverse, with complex operational requirements:

- **Availability—**continuous access to the distributed data warehouse and Web sites
- **Security—**appropriate controls for access and information assurance
- **Data protection—**loss of data is decidedly unacceptable, and backup is critical
- **Data mobility—**data need to be available to the right user, at the right time, in the right place
- **Data purpose—**the same data may have multiple purposes and views
- **Data sharing—**access to all information by all participants
- **Real-time availability—**data must be available at all times in a global setting[15]

IBM, a leading vendor in bioinformatics tools, proposes secure access to data from a growing number of increasingly diverse data sources and the ability to put that data to use quickly; simplified sharing of data and functionality among the diverse applications and tools used in different research areas; easier collaboration internally and externally to turn data into knowledge, as well as the ability to manage and share that knowledge more efficiently; secure storage and easier management of data; faster installation of new applications and integration with valuable existing systems, making research and product development more efficient; and smooth integration of outsourced functions.

### *Additional Considerations*

Computing professionals in bioinformatics will also have to deal with many of the following public issues:

- **Bioethics—**The moral and ethical implications in the application of bioinformatics to genetics. For example, is the manipulation of human cells via genetic engineering contrary to the laws of nature and religion? Cloning is yet another issue.
- **Intellectual property—**The ownership of the human genome is probably the most critical issue. Researchers at universities where a great deal of bioinformatics research is done should clarify intellectual property issues with the university. Ownership of the successful experiments performed “in silico” (via the computer chip) is an unresolved question.
- **Responsibility—**Who is responsible for the results? When errors cause injury or damage, who will be responsible?
- **Access—**Who should have access to the data and for what use? Should law enforcement, insurance companies, HMOs, and employers have access?
- **Privacy—**How will privacy be protected? Who controls the information? How will conformance to laws like HIPAA be enforced?
- **Standards—**In terms of gene therapy, what is normal and what is a disability or disorder?
- **Technology access—**How will the digital divide between those who do and do not have access to expensive technologies be reconciled?
- **Outsourcing—**How will outsourcing affect the field? Given the sensitive nature of research in bioinformatics, what additional legal and intellectual property rights issues will develop?

Advances in the understanding of human genetics and genomics will have important implications for individuals and society. Examination of the ethical, legal, and social implications of genome research is

therefore an integral component of bioinformatics. Collaboration among biological and social scientists, healthcare professionals, historians, legal scholars, ethicists, social workers, and others is essential to the continued debate.[16]

## Conclusions

These are exciting times for bioinformatics and computing, with great career opportunities in developing sophisticated computing tools, including databases and data warehouses, Web-based retrieval and query applications, search engines, analytical and data mining software, knowledge management, and storage applications. The design, implementation, and use of these tools in genomic and related research areas will keep computing professionals busy and productive for a long time. The overall impact of the Human Genome Project can be felt in the need for new types of scientific professionals willing to work in the integrative field of bioinformatics. Not only is knowledge regarding computing crucial, but so is domain knowledge in genomics and the life sciences. The challenge lies in training more individuals who are excited and willing to work in these interdisciplinary areas. Simultaneously, collaboration with other disciplines, including the life sciences and molecular biology, as well as consideration of public and social policy issues will enhance the debate on future applications.

The triad model described here provides for a framework for discussion in the computing field. Additionally, research constructs and designs can be developed to examine the relationships, responsibilities, and roles of the participants in the model. The generalizability and transferability of the lessons learned and technologies to other global collaborative research involving large-scale multi-user problems are envisioned. The promise of hope from genetic studies in diagnosis and treatment of diseases[17] can be fulfilled by the advances in computing technology and its many facets, as well as by addressing the surrounding ethical, public policy, and social issues. The contingent of healthcare workers, providers, recipients, ethicists, sociologists, computing professionals, and scientists will have to fulfill the important role of achieving consensus. Operationally, advancing concepts such as grid, pervasive, and ubiquitous computing offers computational power for the collaborative and on-demand services demanded by bioinformatics. As the healthcare field grapples with the rapid development of information technology and its potential application in reducing overall costs, issues such as privacy and security relating to HIM need to be continually debated. Bioinformatics is only one component in this larger field. This article provides a starting point for discussion of the various public policy issues.

*Someswar Kesh, PhD, is a professor in the Computer Information Systems Department at Central Missouri State University in Warrensburg, MO. Wullianallur Raghupathi, PhD, is an associate professor of Information Systems at Fordham University Graduate School of Business in New York.*

## Notes

1. Available at the National Center for Biotechnology Information's Web site at www.ncbi.nlm.nih.gov.
2. Ibid.
3. Attwood, Teresa K. "Genomics. The Babel of Bioinformatics." *Science* 290, no. 5491 (2000): 471-473.
4. Gibson, Greg and Spencer V. Muse. *A Primer on Genome Science.* Sunderland, MA: Sinauer Associates, Inc., Publishers, 2002.
5. Thornton, Janet M. "From Genome to Function." *Science* 292, no. 5524 (2001): 2095-2097.

6. Hlodan, Oksana. “For Sale: Iceland’s Genetic History.” Available at the American Institute of Biological Sciences’ Web site at www.actionbioscience.org/genomic/hlodan.html.

7. Baxevanis, Andreas. D. and B. F. Francis Ouellette (Editors). *Bioinformatics, 2nd edition.* New York: John Wiley & Sons, Inc., 2001.

8. Westhead, David R., J. Howard Parish, and Richard. M. Thyman. *Bioinformatics.* Oxfordshire, UK: BIOS Scientific Publishers, 2002.

9. Swope, William. C. “Deep Computing for the Life Sciences.” *IBM Systems Journal* 40, no. 2 (2001): 248-262.

10. Mount, D.W. *Bioinformatics, Sequence and Gene Analysis.* New York: Cold Spring Harbor Laboratory Press, 2001.

11. Lesk, A. M. *Introduction to Bioinformatics.* Oxford, UK: Oxford University Press, 2002.

12. Head-Gordon, Teresa and John C. Wooley. “Computational Challenges in Structural and Functional Genomics.” *IBM Systems Journal* 40, no. 2 (2001): 265-291.

13. Westhead, David R., J. Howard Parish, and Richard M. Twyman. *Bioinformatics.*

14. Graham-Rowe, Duncan. “Software Agents Could Tackle Human Genome Data Explosion.” *New Scientist* 179, no. 2407 (2003): 22.

15. Goble, Carole A. et al. “Transparent Access to Multiple Bioinformatics Information Sources.” *IBM Systems Journal* 40, no. 2 (2001): 532-551.

16. Sensen, Christoph W. (Editor). *Essentials of Genomics and Bioinformatics.* Weinheim, Germany: Wiley-VCH Verlag GmbH & Co., 2002.

17. Regalado, Antonio and Leila Abboud. “New Genetics Map to Explore Links to Ailments.” *The Wall Street Journal* October 30, 2002, p. D4.